Mapping the GALEN CORE Model to SNOMED-III: Initial Experiments

PM.Pole MB ChB and AL Rector MD PhD
Medical Informatics Group, Department of Computer Science
University of Manchester, UK

Abstract

As an experiment, a physician without prior training developed a model of two contiguous segments of the SNOMED disease chapter along with related sections from the Topography, Function and Morphology chapters. Effort required, coverage, reasons for difficulties and added value to SNOMED are examined.

INTRODUCTION

The GALEN project is developing re-usable compositional models of medical concepts to support a comprehensive 'Terminology Server' and data entry systems for clinical applications [1-3]. A key goal is to produce a Common Reference Model for medical terminologies which can act as a reference point and interlingua for clinical applications, extending classifications, and converting among coding systems. GALEN aims to bridge the gap between the fine detail required for clinical care and the more abstract categories typically found in the traditional classifications cross-referenced in the UMLS metathesaurus [4], the majority of which are designed primarily for epidemiology, statistical returns or management and remuneration.

The most comprehensive single potential starting source for such a model is the SNOMED International nomenclature [5, 6]. SNOMED is compositional, but it provides only a fixed number of chapters or 'fascicles'- topography, morphology, function, etc. plus a disease chapter which provides specific links to ICD-9 cross referenced to the individual chapters.It is extremely rich and detailed with some cross-references. However it has a number of well recognised limitations.

- Each SNOMED chapter or fascicle is limited to a fixed depth of five or six levels in a single hierarchy.
- SNOMED does not provide explicit semantics for the links between the different elements of a compound term eg whether location, cause, etc.
- It does not provide any general mechanism for classifying compound concepts, and it provides no notion of a canonical form for concepts i.e. - there are many medical concepts which can be represented by several different SNOMED expressions.

One way to address these problems is to model the SNOMED expressions using GALEN's modelling techniques [7], which have unlimited depth, a well-defined semantics, and provide classification and canonical forms for composite concepts automatically.

Major issues in this process concern how much effort is required to do this task manually; how large the models will be; and whether or not they appear to give added value for that effort. This paper reports a preliminary experiment in which one clinician, initially inexperienced in both GALEN and SNOMED, produced a model to represent a contiguous segment of the SNOMED disease chapter.

BACKGROUND

The use of a compositional model as an interlingua or reference model dates back at least to the work of Masarie [8] and has recently been advocated by a position statement and series of papers from the CANON group [9]. Campbell [10] has used SNOMED terms within a representation based on Sowa's Conceptual Graphs [11] and Leveques has used similar ideas in a browsing system for SNOMED [12].

GALEN uses the GALEN Representation and Integration Language (GRAIL), a description logic closely related to frame systems and conceptual graphs but with special features to manage transitive relations such as *part-of* and *cause-of* as well as *is-kind-of* [13, 14] and designed to counter some of the early difficulties encountered in attempts to use a terminological formalism for medical language [15].

Two key features of GRAIL are that it is compositional and generative *i.e.* that new entities can be created by the combination of existing entities and that such composition is sufficiently constrained that only entities representing medically sensible concepts can be generated. The goal is that from a given minimal model or 'basis', it should be possible to generate many different extensions by systematically combining the different basic entities, *e.g.* to produce a classification of cardiology organised by anatomy or function or both.

PURPOSE OF STUDY

The purpose of this study was four-fold:

0195—4210/96/$5.00

Coverage - To see whether the range of concepts covered by a section of SNOMED could be represented in GRAIL , and therefore by extension whether SNOMED could serve a key source and reference point for further development of the CORE model.

Effort required -To obtain preliminary estimates of size of the model and the effort required to develop it by relatively non-specialist staff.

Value to SNOMED - To explore the potential added value to SNOMED of providing a more decomposed and highly structured representation in GRAIL, a topic which will be explored in more detail in a separate paper.

Value of Pre-Modelling fascicles - To confirm that it is more efficient to model the relevant sections of the topography, function, and morphology fascicles of SNOMED before modelling the disease section rather than performing the two tasks in parallel.

The specific objective was taken as representing a complete contiguous segment of a SNOMED Disease chapter along with all of the concepts from the other fascicles cross referenced to it.

METHOD

Selection of Scope of Codes to be Modelled

Cardiology and Respiratory Medicine were chosen as the domain for both experiments because of external demands on the project. Because of the close links between ICD and SNOMED, the corresponding chapters of ICD9 were also mapped. The study was deliberately conducted prior to the construction of the GALEN models for cardiac anatomy and function to test the hypothesis that the two models could be built concurrently. Subsequently the respiratory disease section of SNOMED codes was modelled by first modelling the relevant portions of the topography and function facscicles and then creating the disease concepts for mapping.

Modelling of Underlying Anatomy and Physiology

Initial analysis - The only relevant anatomical term already present was 'Heart', but the model did contain pathological process terms such as inflammation, infection,and ineffective function and acquired lesion terms such as aneurysm, infective lesion and plaque.

Initial examination also indicated that general extensions of the then existing topographical model would be required to cope with junctions and valves. Process terms concerning general diseases were also added where required, *e.g.* Rheumatic Fever was defined as "an inflammatory process which is a late consequence of infection of the pharynx by Streptococcus pyogenes". It was left to the discretion of the clinician to determine what level of detail was appropriate, guided by their intuition of clinical use and potential axes for retrieval such as 'sequellae of Streptococcus infection'.

Although the detail included was greater than that strictly necessary to map the SNOMED codes, the richer descriptions formed a very small part of the exercise. It was felt that the additional detail makes the exercise more realistic as it is expected to be necessary for many other uses of the model.

Code mappings - Codes were then mapped to the concepts that had been created in the expansion of the model, some of the concepts being made as the mapping was created.*eg* Cardiac Aneurysm D3-10500 was identified with the GRAIL concept *Aneurysm which hasLocation Heart.* A number of disease codes refer to combinations of diseases, *e.g.* Rheumatic Fever without heart involvement. In accordance with standard GALEN practice, therefore, all disease codes were mapped to '*ClinicalSituation*' which could involve the presence or absence of one or more disease descriptors, *e.g.*

Clinical situation which shows <
(presence which isExistenceOf RhematicFever)
(absence which isExistenceOf HeartPathology)>

Mapping of the Topography and Morphology Chapters - SNOMED Cardiology Topography and General Morphology Sections were modelled and mapped to explore these as sources for further development of the model. Codes were initially selected on the basis of being cross-referenced within SNOMED to the cardiovascular diseases section but this high-lighted the inadequacy of the cross-referencing as there were many terms which had been used in the disease mapping task but did not have cross-references attached.

Mapping of the Respiratory Section - The Respiratory section was then undertaken, using the respiratory topography and function sections as source for model extension before reference to the disease section. This proved to be a far better approach as most of the terms required in the disease section were then in place and could be combined to express respiratory diseases. Terms which were ommited from the topography section were those which were too

fine detailed to be relevant to disease terms; terms omitted from the function section were those which referred to detailed measurements of respiratory function.

RESULTS

Model Size

	Before CVS	After CVS	Total Added
Concepts	3005	4797	1792
Grammatical Sanctions	340	428	88
Sensible Sanctions	888	2738	1850
Necessary Statements	616	1475	859

Table 1 : Model Size

Table 1 gives the increase in model size for the cardiovascular expansion. The number of concepts shown indicates the number entities required as components in the composite structures used to model the SNOMED terms. They do not equate itemised terms in SNOMED. In GRAIL nothing may be combined into a composite concept until it is sanctioned. 'Grammatical sanctions' are abstract statements permitting queries such as "lesions can be located in anatomical structures". 'Sensible sanctions' are specific statements allowing generation such as "Ventricles may have laterality left and right". 'Necessary statements' represent anatomical facts such as "Heart valves have cusps".

Coverage

SNOMED Cardiology	Codes in Section	Codes Mapped	Percentage Mapped
Topography	206	157	76%
Morphology (Gen)	235	210	89%
Disease	391	357	91%
ICD Cardiology	218	199	91%

Table 2 : C odes Mapped to GRAIL from SNOMED and ICD-9 Cardiology Sections

SNOMED Respiratory	Codes in Section	Codes Mapped	Percentage Mapped
Topography	399	346	87%
Function	337	159	47%
Disease	485	412	85%

Table 3 : Codes Mapped to GRAIL from SNOMED Respiratory Section

Tables 2 and 3 show the coverage achieved within the experiment. The low rate of respiratory function reflects the relatively primitive state of the overall functional model at the time of the experiment.

Codes Not Mapped

Codes which were not mapped fell into groups which were identifiable and similar in both the cardiology and the respiratory sections.

	ICD-9 Cardiology	SNOMED Cardiology	SNOMED Respiratory
NEC	10	1	2
Diseases in Diseases Classified Elsewhere	4	0	0
Oddities	3	8	13
difficult to model in GRAIL	2	7	18
Specialist Terms	0	19	41

Table 4 : Codes not mapped

Not Elsewhere Classified (NEC) - This form is largely confined to ICD-9 but one existed in SNOMED- Other acute rheumatic heart disease Not Elsewhere Classified D3-17460, Acute respiratory distress NEC D2-60250 and Allergic Asthma with stated cause D2-51110. (This last is taken as being a paraphrase of NEC.) Because the meaning of 'not elsewhere classified' depends on the source classification, an exact mapping in GRAIL would require explicit negation of all the possibilities. GRAIL does not support negation. Although there is a GALEN convention for dealing with 'NEC' by creating special elementary entities, these cases were counted as 'not mapped' for this exercise.

Diseases in diseases classified elsewhere As in NEC, definitions which refer to the given coding system are difficult to give an absolute meaning in the GALEN model *e.g.* Acute Pericarditis in diseases classified elsewhere D420.0.

Odd terms - These were terms whose formal meaning was unclear, *e.g.* Obscure cardiomyopathy of Africa(SNOMED D3-20010 ICD D425.2), or imprecise, *e.g.* Anomalous atrioventricular excitation(SNOMED D3-31140 ICD D426.7)or ambiguous, *e.g.* Functional disturbances following cardiac surgery(SNOMED D3-16130 ICD D429.4). This last was considered ambiguous because of the use of 'functional' to mean either 'psychiatric' or 'relating to function' although cross referencing in SNOMED suggested the latter was intended.

Difficult to model in GRAIL - Only two terms could not be modelled because of lack of expressiveness in the GRAIL modelling language, both from ICD-9. These were Kyphoscoliotic heart disease(SNOMED D3-40004 ICD D416.1) and Trifascicular Block(SNOMED D3-33400 ICD D426.54). A descriptive definition of 'Kyphoscoliosis' requires a more sophisticated spatial model than

was available. 'Trifascicular block' requires expressions for numbers, which were deliberately omitted from GRAIL because of concerns over computational complexity and preliminary data suggested that they occurred in less than one percent of all terms to be represented. Any such terms, can of course, be added as primitives with partial descriptions using 'necessary statements', but will not be fully classified by the GRAIL classifier.

Specialist terms - The section on cardiac dysrhythmias required specialist knowledge to model which was beyond the clinical expertise of the clinician working on the exercise.

Effort Required

Initial training - 6 months

Cardiovascular - 2 months for 724 codes (20 codes per day)

Respiratory - 1.5 months for 917 codes (30 codes per day)

Potential Added Value for SNOMED

A specific example of the potential for re-organisation of ICD or SNOMED using GRAIL would be as a tool by which Rheumatic fever (SNOMED D3-17100 ICD D39) could be taken out of Cardiology and become a systemic disease which has diverse pathology, including cardiac pathology, as a consequence.

This would overcome the problem of Rheumatic Fever without heart involvement (SNOMED D3-17110 ICD D390) appearing within the Cardiovascular disease section, and similarly Rheumatic Chorea (SNOMED D3-17116 ICD D392.9) which should really be classified as a kind of Movement Disorder (SNOMED DA.21100) or Chorea (ICD D333.5) and specialised by being a consequence of Rheumatic Fever, and Rheumatic Joint disease (SNOMED D3-17111 Not in ICD) which should be a kind of Arthropathy (SNOMED D1-20000 ICD D71). Using GRAIL these could all be classified within their appropriate disease section, but could also be collected together as diseases which are a consequence of Rheumatic Fever.

DISCUSSION

Achievement of Goals

Coverage - It is possible to represent the bulk of SNOMED terms using the GRAIL formalism. Many of the terms which could not be mapped might be considered artefacts of the coding system, such as 'Not elsewhere classified'.

Value to SNOMED - The additional structure of GRAIL provides a means of organising terms by meaning and was not constrained to a simple hierarchy. The potential for use is illustrated in the above discussion and further experiments are planned.

Effort Required - This experience projects to a requirement of 1,000 person days (4 years) to map the disease section (30,000 terms) in SNOMED. This suggests manual modelling may be too labour intensive to be justifiable. There are opportunities for the application of a semi-automated approaches using automatic parsing of rubrics and generation of GRAIL concepts from the parsed terms. These methods are under development currently.

The six month training period for the current worker has now been reduced to three months with improved training material.

There is a very steep learning curve during which many new skills must be acquired and new understanding gained. The method of analysis and subsequent construction of new terms requires a new way of viewing the world. Once established, it is difficult to return to 'not-knowing' and therefore difficult to explain to a new-comer. This has been answered in the group by the teacher of the new-comer being the newest in the group and closest to the learning process themselves. This reduces the difficulty of communication but increases the risk of evolutionary drift from the original modelling method, although all models are reviewed and revised by a more experienced modeller before release.

SNOMED as a source for further modelling

The topography, morphology and function sections of SNOMED are a useful source for the basic terms on which medicine depends. As a source for expansion of the model for use in a clinical context SNOMED is useful though filtering by a clinician may be needed to exclude excess details and pathology bias. Experiments are underway to use linguistic methods for a first pass translation, and will be reported subsequently.

GRAIL's inability to express numbers (cardinality) was a problem in a real but small number of cases (2 out of 391). This is consistent with preliminary studies during the design phase suggesting that a lack of detailed cardinality constraints would affect less than 1% of terms. Other limitations within the formalism did not present serious problems within this corpus.

Having established the feasibility and scale of the work the next stage is the creation of a re-usable model with broad scope and multiple

applications suitable for a general terminology server. Other work has established the value of such a model for supporting structured data entry [16, 17] and work is proceeding on using such models to mediate between different medical record architectures and between medical records and decision support systems [18] .

Acknowledgements

This work was supported in part by Grant 2012 from the EU AIM Initiative. It includes work from many partners in the GALEN consortium.

References

1. Rector A, Nowlan W, Glowinski A. Goals for Concept Representation in the GALEN project. 17th Annual Symposium on Computer Applications in Medical Care (SCAMC-93). McGraw Hill, 1993: 414-418.

2. Nowlan W, Rector A, Rush T, Solomon W. From Terminology to Terminology Services. 18th Annual Symposium on Computer Applications in Medical Care (SCAMC-94). Washington, DC, 1994: 150-154.

3. Rector A, Solomon W, Nowlan W, Rush T. A Terminology Server for Medical Language and Medical Information Systems. *Methods of Information in Medicine* ,1995:34:147-157.

4. Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. In: van Bemmel J, ed. 1993 Yearbook of Medical Informatics. Amsterdam: Intermational Medical Informatics Association, 1993: 41-53.

5. Côté R, Rothwell D. *SNOMED-International*.Chicago: College of American Pathologists, 1993

6. Chute C, et al. The content and coverage of clinical classifications. *JAMIA* 1996;3:224-233.

7. Rector A. Coordinating taxonomies: Key to re-usable concept representations. In: Barahona P, Stefanelli M, Wyatt J, (ed). Fifth conference on Artificial Intelligence in Medicine Europe (AIME '95). Pavia, Italy: Springer, 1995: 17-28.

8. Masarie F, Miller R, Bouhaddou O, Giuse N, Warner H. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. *Computers in Biomedical Research* 1991:24(4):379-400.

9. Evans DA, Cimino J, Hersh WR, Huff SM, Bell DS, The Canon Group. Position Statement: Towards a Medical Concept Representation Language. *Journal of the American Medical Informatics Association* 1994:1(3):207-217.

10. Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. *JAMIA* 1994:1(3):218-232.

11. Sowa J. *Conceptual Structures: Knowledge Representation in Mind and Machine*.New York: John Wiley & Sons, 1985

12. Lévesque Y, LeBlank Ar, Maksud M. MD Concept: A model for integrating medical knowledge. In: Ozboldt JG, (ed). Eighteenth Annual Symposium on Computer Applications in Medical Care. Washington DC: Hanley & Belfus, Inc, Philadelphia PA, 1994: 253-256.

13. Rector AL, Nowlan W. A Reusable Application Independent Model of Medical Terminology: GALEN's GRAIL. KR-94. Berline: Morgan Kaufmann, 1994: in press.

14. Rector A, Bechhofer S, Goble C, Horrocks I, Nowlan W, Solomon W. The GRAIL Concept Modelling Language for Medical Terminology. *Artificial Intelligence in Medicine* 1996(in press).

15. Doyle J, Patil R. Two theses of knowledge representation: Language restrictions, taxonomic classification and the utility of representation services. *Artificial Intelligence* 1991:48:261-297.

16. Kirby J, Rector AL. The PEN&PAD Data Entry System: From prototype to practical system. AMIA Fall Symposium. Washington DC, 1996: in press.

17. Nowlan W, Rector A. Medical Knowledge Representation and Predictive Data Entry. In: Stefanelli M, Hasman A, Fiesch M, Talmon J, (ed). Artificial Intelligence in Medicine Europe (AIME-91). Maastricht: Springer-Verlag, 1991: 105-116.

18. Rector A. Compositional models of medical concepts: towards re-usable application-independent medical terminologies. In: Barahona P, Christensen J, ed. Knowledge and Decisions in Health Telematics. IOS Press, 1994: 133-142.